Curso Práctico de Bioestadística Con Herramientas De Excel

Fabrizio Marcillo Morla MBA

barcillo@gmail.com
(593-9) 4194239

Fabrizio Marcillo Morla

- Guayaquil, 1966.
- BSc. Acuicultura. (ESPOL 1991).
- Magister en Administración de Empresas. (ESPOL, 1996).
- Profesor ESPOL desde el 2001.
- 20 años experiencia profesional:
  - Producción.
  - Administración.
  - Finanzas.
  - Investigación.
  - Consultorías.

Otras Publicaciones del mismo autor en Repositorio ESPOL

Capitulo 3

Estadistica Descriptiva

Estadistica Predictiva

- Datos estadísticos, obtenidos de muestras, experimentos o cualquier colección de mediciones, a menudo son tan numerosos que carecen de utilidad a menos que sean condensados o reducidos a una forma más adecuada.
- En esta sección nos ocuparemos del agrupamiento de datos, así como de ciertos estadísticos o medidas que representarán el significado general de nuestros datos.

Distribucion de Frecuencias

- Operación en que dividimos conjunto datos en un número de clases apropiadas, mostrando también el número de elementos en cada clase.
- Se pierde algo de información, pero ganamos claridad.
- 1ª etapa decidir cuántas clases y elegir límites
- Número clases dependerá número y rango datos
- Matemáticamente, # intervalos (k) :

$$k = 1 + \frac{10}{3} \ln n$$

- Hay que ver qué tan bien representa esto a los datos.
- En general se recomienda k entre 5 y 15.

Distribucion de Frecuencias

- Intervalo de representación: intervalo donde se representan los datos.
- Intervalo real: verdaderos límites intervalo. Punto medio entre límites dos int. representación consecutivos
- Marca de clase: punto medio intervalo de representación.
- Frecuencia: Cantidad ocurrencias de datos dentro de un intervalo de representación.
- Frecuencia relativa: Relación entre la frecuencia de un intervalo y la frecuencia total expresada en porcentaje.
- Frecuencia acumulada y acumulada relativa son suma de número ocurrencias o porcentajes de todos los intervalos menores o iguales al presente.

Histograma

- Rectángulos representan frecuencias de clase
- Bases se extienden en las fronteras de los intervalos reales.
- Marcas de clase situadas en la mitad del rango del rectángulo.
- Podemos usar para frecuencia o f. relativa, pero no para f. acumulada o acumulada relativa.

Diagrama de Barras

- Similares a los histogramas
- Alturas y no áreas representan frecuencias
- No se pretende fijar ninguna escala horizontal continua
  - El ancho de las barras no interesa.
- Se pueden graficar tanto f. absolutas o relativas, así como las acumuladas

Poligonos de Frecuencia

- Frecuencias de clases graficadas sobre marcas de clase y unidas mediante líneas rectas.
- Agregamos valores correspondientes a cero en los puntos límites de la distribución.
- Podemos urepresentar indistintamente las frecuencias netas o acumuladas
- Para acumuladas, en vez de usar marcas de clase como abscisas utilizamos el límite superior del intervalo real de frecuencia.

Graficos de Sectores

- Tambien llamado Grafico de Pastel
- Para frecuencias relativas
- Corresponde a un círculo dividido en varios sectores, correspondiendo cada uno a un intervalo
- Area de cada sector es proporcional a la frecuencia relativa.

Estimación de Parámetros

- Sirve para describir poblaciones.
  - Ej: resultados de una prueba.
- Estimación puntual: elegir un estadístico calculado a partir de datos muestreales, respecto al cual tenemos alguna esperanza o seguridad de que esté "razonablemente cerca" del parámetro que ha de estimar.
- Estimación puntual no es mas que calcular un estadístico, y decir que este estadístico esta "razonablemente cerca" del parámetro poblacional.

Estimadores

- Para poblaciones normales, el estimador más eficiente de μ es el promedio ($\bar{x}$).
- Para la varianza poblacional, el estimador insesgado más eficiente es la varianza muestreal.
- Rango muestreal R, se puede sacar estimador insesgado de σ.
  - Relación R/d2 para n$\leq$ 5 mas eficiente que s
  - Valores de d2 para distintos valores de n:

n	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
d_2	1.128	1.693	2.059	2.326	2.534	2.704	2.847	2.970	3.078

Estimadores

- Para proporciones, estimador insesgado más eficiente de parámetro proporción poblacional (p) es estadístico proporción muestreal (x/n):

$$x / n = \frac{x}{n}$$

- X: # observaciones con un caracter determinado y n es número total de observaciones (x + ¬x).

Estimación Por Intervalos

- Cuando usamos estadístico para estimar parámetro, $P(\theta_0=\theta)$ prácticamente nula.
- Es conveniente acompañar estimación puntual con el error de estimación que probablemente tenemos
- Estimación por intervalos:
  - Probabilidad que parámetro esté dentro ese intervalo.
- Forma de estimar parámetros depende del parámetro y del tipo de muestreo.
- Probabilidades varian por tipo de muestreo.

Que Tipo Muestreo y n Uso?

- Depende de cuanta información se quiera y se pueda conseguir.
- Especificar límite para error de estimación:
  - θ y θ_0 difieran en cantidad menor que Δ: **$E \leq \Delta$**.
- Especificar probabilidad ($1-\alpha$):
  - % veces que al muestrear repetidamente la población, error de estimación sea menor a Δ:

$$P(E \leq \Delta) = 1-\alpha$$

- Luego elegir método con mayor precision a menor costo.

Error y Tamaño Muestra

- Dos factores influyen en la cantidad de información contenida en una muestra.
- Tamaño de la muestra
- Variación entre individuos de población
- Si variación es variable dependiente, puede ser controlada por método de muestreo.
- Para mismo tamaño muestra fija, considerar varios muestreos:
  - Muestreo cuesta plata
  - Diseño que estime mas preciso con menor n da ahorro en costo experimentado.

Muestreo Totalmente Aleatorio

- Muestreo irrestricto al azar
- Seleccionar un muestreo de n individuos de tal forma que cada muestra de tamaño n tenga la misma oportunidad de ser seleccionada.
- Muestra se la llama muestra totalmente aleatoria
- Igual de bueno como otros siempre y cuando:
  - Todos individuos población sean similares en cuanto a información que nos interese
  - No exista otra variable que no permita separarla en grupos distintos entre ellos, pero mas homogenos dentro de ellos que la población original.

Estimación de Medias

- Para estimar μ usamos el promedio $\bar{x}$:

$$\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Error de estimación para poblaciones infinitas o muy grandes respecto a la muestra será:

$$E = Z_{(\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}}$$

- Poblaciones finitas, o cuando muestra es alto porcentaje de población:

$$E = Z_{(\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}} \sqrt{\left(\frac{N-n}{N}\right)}$$

Pequeñas Muestras

- Las fórmulas antes descritas funcionan bien cuando se conoce σ^2, o n>30,
- De lo contrario, siempre y cuando podamos suponer razonablemente que estamos muestreando de una población Normal,debemos estimar usando t:

$$E = t_{(\frac{\alpha}{2})} \cdot \frac{s}{\sqrt{n}}$$

- Para un porcentaje de confianza de 100 x (1-α) y para ν = n-1 grados de libertad.

Estimación de Varianzas

- Para estimar la varianza poblacional utilizaremos el estadístico varianza muestreal:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{(n-1)}$$

- El intervalo de confianza vendrá dado por:

$$\frac{(n-1)s^2}{\chi^2_{(\alpha/2)}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}}$$

Estimación de Proporciones

- El estimador para la proporción poblacional p vendrá dado por la proporción muestreal x/n:

$$p = x / n = \frac{x}{n}$$

- Y su error de estimación por:

$$Z_{\alpha/2} \sqrt{\frac{\frac{x}{n}(1 - \frac{x}{n})}{n-1} \left(\frac{N-n}{N} \right)}$$

Tamaño de la Muestra

- Para determinar el tamaño de la muestra utilizaremos la siguiente formula para medias:

$$n = \left[\frac{Z_{(\frac{\alpha}{2})}\sigma}{\Delta} \right]^2$$

- La cual no es mas que la fórmula del error despejada, y en donde n es el tamaño de la muestra, σ es la varianza y Δ el máximo error que estamos dispuestos a aceptar.

Tamaño de la Muestra

- Para proporciones utilizaremos:

$$n = \frac{N(p)(1-p)}{(N-1)\Delta - p(1-p)}$$

- Lógico que estas fórmulas debemos usar antes de muestreo: desconoceremos σ y p.
  - Estos valores se pueden obtener de poblaciones similares, muestreos anteriores a dicha población, o un muestreo de prueba.
  - Para proporciones podemos remplazar p por 0.5 para obtener un tamaño de muestra conservador.

Muestreo Aleatorio Estratificado

- Obtenida mediante separación de elementos de población en grupos que no traslapen, llamados estratos
- Selección posterior de muestra aleatoria simple dentro de cada estrato.
- Objetivo al diseñar muestreo: maximizar información obtenida a un costo dado. Este tipo de muestreo puede ser mas eficiente que el totalmente aleatorio bajo ciertas condiciones:
  - Seleccionar estratos donde información va a ser mas homogénea que en la población en general.
  - Necesitamos saber tamaño de estratos.

Muestreo Aleatorio Estratificado

- Obtenida mediante separación de elementos de población en grupos que no traslapen, llamados estratos
- Selección posterior de muestra aleatoria simple dentro de cada estrato.
- Objetivo al diseñar muestreo: maximizar información obtenida a un costo dado. Este tipo de muestreo puede ser mas eficiente que el totalmente aleatorio bajo ciertas condiciones:
  - Seleccionar estratos donde información va a ser mas homogénea que en la población en general.
  - Necesitamos saber tamaño de estratos.

Muestreo Aleatorio Estratificado

- Especificar claramente los estratos.
  - C/individuo esta en uno y solo un estrato apropiado
- Seleccionar una muestra totalmente aleatoria en cada estrato mediante la técnica ya descrita
- Muestras seleccionadas en cada estrato seran independientes.
  - Muestras seleccionadas en un estrato no dependan de las seleccionadas en otro

Definiciones

- **Número de estratos:** L
- **Numero de individuos en estrato i:** Ni
- **Número de individuos en población:** $N = \Sigma Ni$
- **Tamaño de la muestra en el estrato i:** ni
- **Media del estrato i:** μi
- **Media de la población:** μ
- **Variaza del estrato i:** $\sigma^2 i$
- **Varianza de la Población:** σ^2
- **Total del estrato i:** τ i
- **Total Poblacional:** τ

Estimación μ, σ^2

- El estimador de μ es x_{st}, st indica muestreo aleatorio estratificado:

$$x_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{x}_i$$

  - Bastante parecido a promedio ponderado.
- y el límite para el error de estimación E :

$$E = Z_{\alpha/2}\sqrt{\frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)}$$

- Estimador de la varianza de x_{st} será:

$$\sigma^2 = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)$$

Estimación p

- Para proporciones, el estimador de la proporción poblacional p vendrá dado por:

$$p_{st} = \frac{1}{N} \sum_{i=1}^{L} N_i p_i$$

- Y los límites para el error de estimación por:

$$E = Z_{\alpha/2} \sqrt{\frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \left(\frac{pq}{n_i - 1} \right)}$$

Grafico de Intervalos

Curtosis y Skewness

- Comparación con distribución es normal.
- Curtosis:
  - Elevación o achatamiento comparada normal.
  - Positiva: relativamente elevada
  - Negativa: relativamente plana
  - = CURTOSIS(rango) o Herramientas Analisis
- Skewness (coeficiente Asimetría, Sesgo)
  - Asimetría respecto a su media
  - Positiva: Sesgo hacia derecha
  - Negativa: Sesgo Izquierda
  - =COEFICIENTE.ASIMETRIA(Rango)
  - o Herramientas Analisis

$$b_2 = \frac{nx\sum(x_i - \bar{x})^4}{(\sum(x_i - \bar{x})^2)^2}$$

$$b_1 = \frac{\sqrt{n}x\sum(x_i - \bar{x})^3}{(\sum(x_i - \bar{x})^2)^{3/2}}$$

Intervalos Confianza en Excel

- Ejercicio08 - Estadistica Descriptiva.xlsx
- =INTERVALO.CONFIANZA(α,σ,n)
  - Da el intervalo de confianza para la media cuando se conoce σ o n>30. Usa $Z_{(\alpha/2)}$
- Herramientas de Análisis / Estadisticas descriptivas:
  - Da el intervalo de confianza para la media cuando se desconoce σ o n<30. Usa $t_{(\alpha/2)}$
  - Recordar que DISTR.T.INV usa 2 colas

Regresion Lineal

- Fijamos valores variable independiente (x), y observamos variable dependiente (y) de ésta.
- Lograr ecuación para describir comportamiento y relacionado con x, dentro rango específico.

$$y = a + bx$$

- Análisis correlación mide, para c/ muestra x y y.
- Grafica pares para ver relaciones entre ellos.
- Calcula algunos estadísticos para determinar la fuerza de la relación
  - Regresión para experimentos reales
  - Correlación para estudios ex post facto
- Puede ser usada como comparativa o predictiva.

Diagrama Dispersión

- Gráfico en el cual van a estar representados, mediante puntos, los valores de nuestros pares de variables (x,y).
- Sirve para darnos una idea visual del tipo de relación que existe entre ambas variables, y debe de ser hecho antes de iniciar cualquier cálculo para evitar trabajos innecesarios
- Excel Grafico dispersión tiene herramientas para evaluación interactiva de correlación.

Mínimos Cuadrados

- Recta donde cuadrados de diferencias entre puntos experimentales (x,y) y puntos calculados (x',y') sea mínima.
- y = a + bx
  - a: intersección de la recta con el eje Y
  - B: pendiente de la recta.

$$b = \frac{\sum xy - \frac{\sum\sum x \sum\sum y}{N}}{\sum\sum x^2 - \frac{(\sum\sum x)^2}{N}}$$

$$a = \frac{\sum\sum y}{N} - b\frac{\sum\sum x}{N}$$

- a =INTERSECCION.EJE(rango Y,rango X)
- b =PENDIENTE(rango Y,rango X)
- Herramientas de Analisis

Coeficiente Determinación

- r^2: proporción de variación en variable y que puede ser atribuida a una regresión lineal con respecto a la variable x:

$$r^2 = \left(\frac{N\sum xy - (\sum x \sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}} \right)^2$$

- Raíz cuadrada positiva (r): coeficiente de correlación de Pearson; estimador parámetro coeficiente de correlación poblacional ρ.
- Eta cuadrado (η^2): relación entre SCT y SC Total del ANOVA. Representa máxima variación total que puede ser atribuida a cualquier regresión de y con respecto de x

Regresiones No Lineales

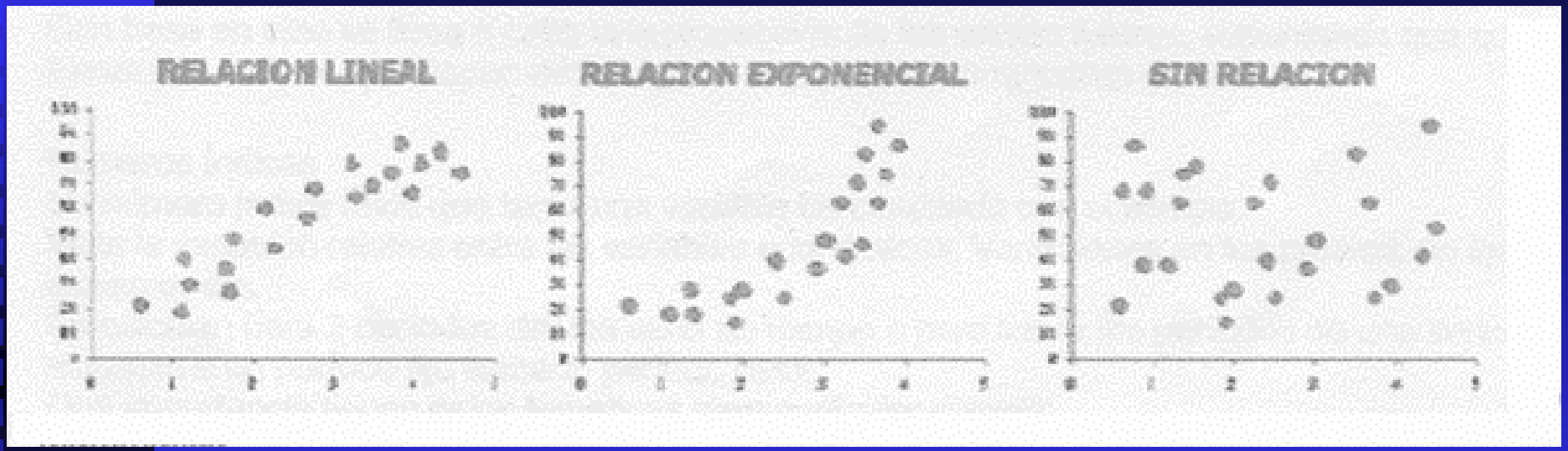
- Existen otros tipos relaciones posibles entre x y y
- Crecimiento poblacional común regresión exponencial:

$$y = ab^x$$

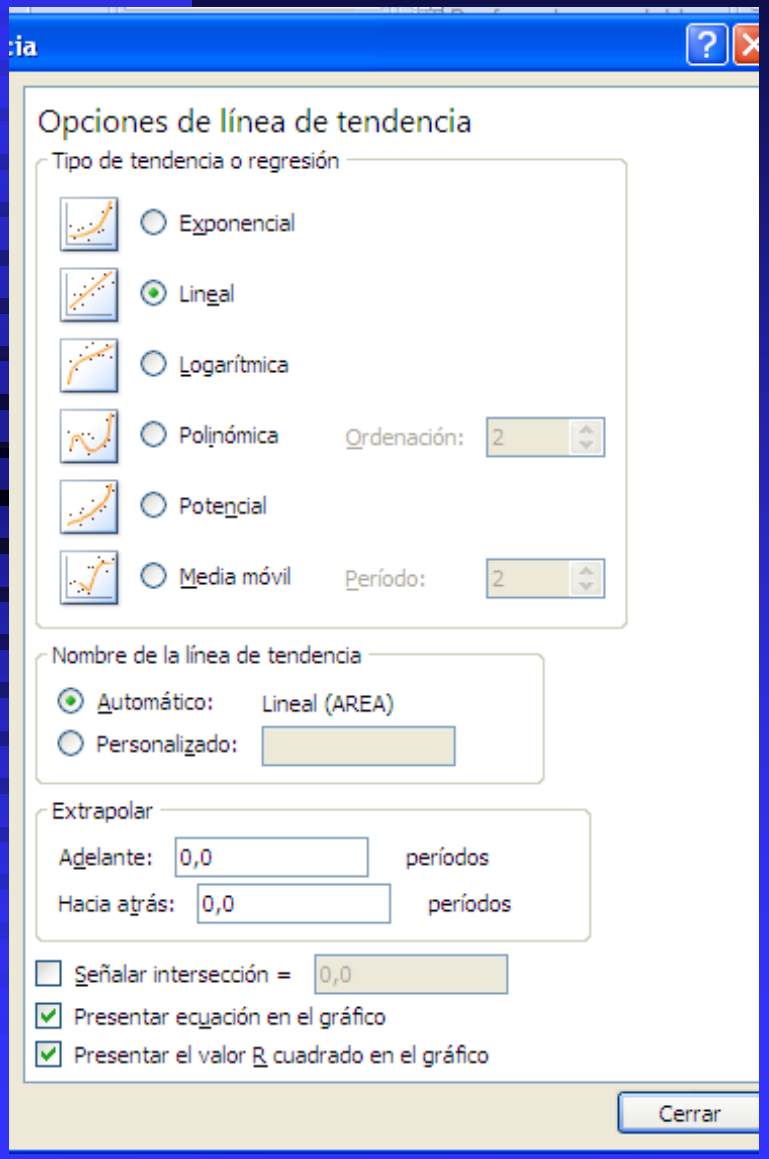
  - a : "índice de Falton“
  - B: índice de crecimiento relativo.
- Grafico en papel semilogarítmico da una línea recta.
- Datos se linealizan con:

$$\log y = \log a + x \log b$$

- Luego es un caso de regresión lineal.


RELACION LINEAL
RELACION EXPONENCIAL
SIN RELACION

Regresiones No Lineales



- Hay otros casos regresiones no lineales y mayoría se linealiza de misma forma.
- Excel presenta opción de visualizar previamente alguns tipos de regresiones visualmente y calcular su ecuación y r2 mediante la opción Formato de Linea de tendencia en los graficos de dispersión.